BEOWULF CLUSTER

EDGE AI

Hoe een Beowulf Cluster opzeten.

Bavo Debraekeleer

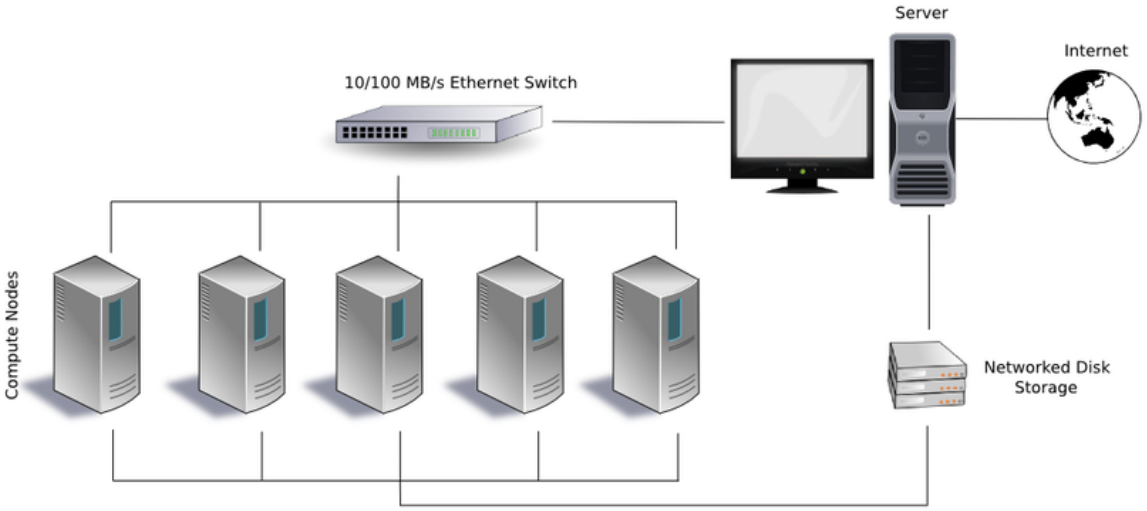
ARTESIS PLANTIJN
HOGESCHOOL ANTWERPEN

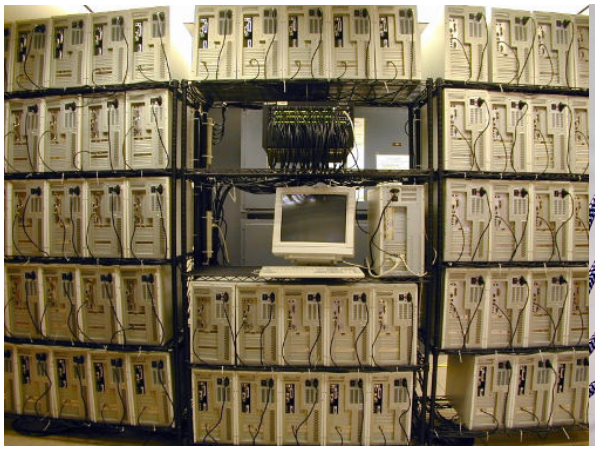
Deliverable

Een verslag over hoe een beowulf cluster kan worden opgezet. Er is geen template of vereiste vorm die je dient te volgen. De vorm van het verslag is vrij, zolang het de architectuur, de achterliggende technologie en de werkwijze duidelijk beschrijft.

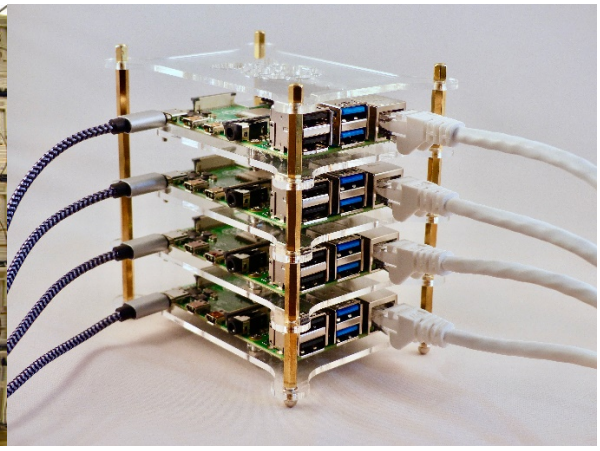
Architectuur

Een Beowulf cluster bestaat uit schaalbare cluster van vrij verkrijgbare computer hardware.
Deze PC's of workstations zijn verbonden op een privaat netwerk (met ethernet switch) en draaien een open source software infrastructure, meestal een Linux distributie.





Beowulf bestaande uit 64 workstations (bron)



Beowulf betsaande uit 4 Raspberry Pi's (bron)

De nadruk bij een Beowulf cluster ligt op snelheid. Het geheel wordt gezien als één systeem die processor kracht en geheugen deelt. Elke afzonderlijke computer binnen de cluster noemen we een Node. Bij het uitvoeren van een applicatie worden de opdrachten opgesplitst naar elke afzonderlijke Node die deze parallel uitvoeren. Zo wordt het resultaat zoveel keer sneller uitgevoerd als dat er Nodes zijn in de cluster.

De computers hoeven ook niet allemaal identiek te zijn. Je kan allerlei soorten mengen. Echter hoe meer variatie hoe moeilijker het wordt om een applicatie optimaal op te splitsen over de cluster.

Om opslag te vereenvoudigen kan een Network File System opgesteld worden en gemount worden in alle nodes. Zo worden veranderingen in deze directory automatisch toegepast op alle nodes. Zonder dit moet je de applicatie telkens kopiëren naar elke node. Op deze manier moet de applicatie op één node gezet worden en wordt dit automatisch gespiegeld op alle andere.

Achterliggende Technologie

Een Beowulf cluster bestaat niet uit strikt bepaalde onderdelen en kunnen op meerdere manieren opgebouwd worden. Veel gebruikte technologieën zijn:

- PVM of Parallel Virtual Machine is een programmabibliotheek die toelaat om computers verbonden in hetzelfde netwerk te gebruiken als één parallelle computer. Hiermee kan je programma's schrijven die parallelle message-passing kunnen verichten, geschreven in Fortran en C. Was de standard tot MPI uit kwam.
- MPI of Message Passing Interface is de nieuwere standard voor portable message-passing parallel programs, gestandaardiseerd door de MPI Forum.
- LAM
- De Linux kernel
- De channel-bonding patch voor de Linux kernel. Deze laat toe meerdere Ethernet interfaces te "binden" zodat je één snellere "virtuele" Ethernet interface bekomt.
- De global pid space patch voor de Linux kernel. Deze laat toe om alle processen in de Beowulf cluster te zien met "ps" en deze ook te elimineren.
- DIPC dat toe laat om "sysv shared memory en semaphores" en message queues transparent te gebruiken over de gehele cluster.

Werkwijze

Wat heb je nodig

- Minstens twee computers
- Internet toegang via Wifi
- Switch en Ethernet kabels

Overweging: denk na over de stroomvoorziening. Voor single board computers zoals de Raspberry Pi kan het interessant zijn een lader te hebben die meerdere uitgangen heeft. Dit is efficiënter, goedkoper en handiger.

Opstelling en installatie

Een cluster werkt met communicatie tussen de Nodes. Een "hoofd" node heeft de leiding over de andere "werker" nodes. De hoofd node zegt wat de werkers moeten doen en vraagt voor rapportering. Hiervoor is het beste dat de cluster zijn eigen lokale netwerk heeft (LAN) met Ethernet kabels. Zo wordt de werking niet in de weg gezeten door ander netwerk trafiek. Om met elke node afzonderlijk te verbinden kan je Wifi gebruiken met internet toegang. Zo blijft de Ethernet poort beschikbaar voor de cluster en hebben de computers internet toegang voor installatie en updates.

1. Installeer een Linux distributie met GCC (GNU Compiler Collection) op de computers.
2. Stel Wifi in voor internet verbinding ("sudo raspi-config" bij RPi).
3. Stel de host names in als "node1", "node2", …

4 Updaten “sudo apt -y update && sudo apt -y upgrade”

De “backbone” netwerk configuratie

De Ethernet link tussen de nodes van de cluster noemt de “backbone”. Hiervoor gebruiken we het 10.0.0.0 subnet. Voor elke node afzonderlijk moet deze ingesteld worden met een statisch IP adres.

5 Ken per node een statisch IP adres toe, vb 10.0.0.1 voor node1, 10.0.0.2 voor node2.
Bij RPi’s doe je dit door vanonder in de file “sudo nano /etc/dhcpcd.conf” het volgende toe te voegen: “interface eth0
static ip_address=10.0.0.1/24”

6 Voeg op elke node ook alle IP adressen en host names toe aan de “/etc/hosts” file: “10.0.0.1 node1” …

SSH instellen

De werker nodes moeten elk kunnen praten met de hoofd node zonder een wachtwoord nodig te hebben. Hiervoor gebruiken we SSH keys. Bij Raspbian voor RPi’s gaat dit out-of-the-box. Bij andere Linux distributies kan het dat je eerst nog ssh-server moet installeren. Dit kan met het commando “sudo apt--get install openssh-server”. Doe dit op elke node.

Optioneel: Het wordt aangeraden, maar is niet noodzakelijk, op elke node een nieuwe user aan te maken met de naam “mpiuser”. De wachtwoorden zijn liefst overal gelijk. Geef deze gebruiker administrator rechten en log vervolgens als deze gebruiker in.

7 Genereer een SSH key op elke node: “ssh-keygen -t dsa” en druk Enter bij elke vraag.
8 Kopieer nu op elk werker node de key naar de hoofd node: “ssh-copy-id 10.0.0.1”
9 En kopieer op de hoofd node de key naar alle werker nodes: “ssh-copy-id 10.0.0.2” …

MPI installeren in instellen

MPI (Message Passing Interface) is het protocol dat het meest gebruikt wordt voor Beowulf cluster. Het laat computers toe taken te delegeren tussen elkaar, te rapporteren en te antwoorden met de resultaten. We gebruiken hier MPI in combinatie met Python.

10 Installatie MPI en Python bindings op elke node: “sudo apt install mpich python3-mpi4py”
11 Test of MPI werkt: “mpiexec -n 1 hostname” (echo host name terug)

Applicaties draaien

12 Processen starten: “mpiexec -n <nr of nodes> --hosts <ip node 1>,<ip node 2> hostname

Voorbeeld op één node: “mpiexec -n 1 python3 prime.py 1000”
Zelfde op meerdere nodes: “mpiexec -n 4 --host 10.0.0.1,10.0.0.2,10.0.0.3,10.0.0.4 python3 prime.py 100000”

Om een applicatie te draaien op meerdere nodes moet deze op elke node staan op dezelfde plaats. Dit kan zoals reeds besproken met netwerk opslag vereenvoudigd worden, maar als je dit niet hebt kan je volgend commando gebruiken: “scp ~/prime.py 10.0.0.x:” (secure copy, vervang x door de node nr)

Elke node krijgt een rang of uniek ID toegewezen. De hoofd node is altijd 0.

Wanneer de resultaten binnen zijn verzameld de hoofd node alles en rapporteert de resultaten.

Bronnen

https://beowulf.org/overview/faq.html

https://www.linux.com/training-tutorials/building-beowulf-cluster-just-13-steps/

https://magpi.raspberrypi.com/articles/build-a-raspberry-pi-cluster-computer